Fall 12-16-2020

Implicit Biases and Byrne's Epistemic Rule for Self-knowledge of Beliefs

Hanyu Liu

Implicit Biases and Byrne's Epistemic Rule for Self-knowledge of Beliefs

by

Hanyu Liu

Under the Direction of Neil Van Leeuwen, PhD

ABSTRACT

Alex Byrne makes a normative claim in his book Transparency And Self-Knowledge, that the epistemic rule for self-knowledge of belief BEL (if P, believe that you believe that P), is a good rule. In this thesis, by utilizing both philosophical writings and empirical studies in psychology I reject Byrne's claim that BEL is a good rule. More specifically, I argue that many cases of implicit biases are essentially beliefs, because they share many characteristics that are paradigmatic to beliefs. Then I argue that applying BEL, as a method of obtaining self-knowledge of beliefs, not only fails to discover many of our implicit biases, but also gives wrong verdict for many of our implicit biases. Finally, I conclude that BEL is not a good rule, and any good method for obtaining self-knowledge of belief should not ignore the importance of observing one's own behaviors.

INDEX WORDS: philosophy of mind, epistemology, self-knowledge, transparency, implicit bias.

Implicit Biases and Byrne's Epistemic Rule for Self-knowledge of Beliefs

by

Hanyu Liu

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

Implicit Biases and Byrne's Epistemic Rule for Self-knowledge of Beliefs

by

Hanyu Liu

Committee Chair: Neil Van Leeuwen

Committee: Daniel A. Weiskopf

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2020

DEDICATION

This thesis is for my mother and my father. I am also grateful for my friends: Botian, De, Kerong, Mingzhu, Rachel, Yuchen Li and Yuchen Liang.

ACKNOWLEDGEMENTS

I am grateful for my thesis advisor and the committee chair, Dr. Van Leeuwen, without whose suggestions and supervision this thesis cannot be completed. I'm also grateful for Dr. Weiskopf and Nahmias' great advices.

TABLE OF CONTENTS

1 INTRODUCTION

Attaining self-knowledge of beliefs, that is, knowing what beliefs we have, is important. It enables us to know more about ourselves, discover potential inconsistencies between our beliefs, and consequently help us regulate ourselves and our relationships with others. So how do we attain self-knowledge of beliefs?

Alex Byrne (2018) offers a theory.[1] He observes that our beliefs are *transparent*, namely, we gain self-knowledge of our beliefs by merely investigating the world: if I see an apple on the table, then I could form the belief that I *believe* that there is an apple on the table. So Byrne provides a descriptive theory of self-knowledge of beliefs that captures transparency: we acquire self-knowledge of beliefs by following or trying to follow an epistemic rule named BEL.[2]

BEL: If *p*, believe that you believe that *p* (Byrne, 103).

To follow BEL, in Byrne's definition, one first knows that *p* is the case, and then believes that one believes that *p*. On the other hand, to *try* to follow BEL, in Byrne's definition, is first to *believe* that *p* is the case, and then believes that one believes that *p* (107). The theory seems intuitive: if someone asks you whether you think it is raining now, you would answer this question by checking whether it is raining—you would look out of the window or check if there is a sound of raindrops. If you recognize that it is not raining, you would then believe that you believe that it is not raining.

After putting forward this descriptive theory, Byrne further argues for a normative claim, that BEL is a good rule in the sense that it is *strongly self-verifying*: if one tries to follow the rule, the

[1] Alex Byrne's overall project in the book is to construct a general theory of self-knowledge that includes the self-knowledge of a wide range of mental states. Here I'm only concerned with his theory of the self-knowledge of beliefs.
[2] An epistemic rule is an inference rule that we implicitly follow to extend our knowledge.

resulting belief is guaranteed to be true.[3] He argues that, when one is trying to follow the rule, one would think that *p* is true, that is, one believes that *p*—and "one believes that *p*" is exactly the content of the resulting belief. Therefore the first state of belief automatically makes the consequent, the second-order belief that one has this belief *p*, true (107).[4]

Byrne's account is attractive, because if we do have this method of self-knowledge of beliefs and it is a good method, then it is easy to know about your own belief about *p*: you simply check if *p* is true. For instance, suppose you are making an important decision about whether to change a job and you ask yourself, do I think my current job is good? By Byrne's account, you could just check whether this job is good: it has a good salary, the workload is not high, and it offers long vacations, so it is a good job. Then by following BEL, you conclude that you think the job is good.

However, what if your colleague heard about your reasoning and is shocked: "No, you cannot actually think that it's a good job! You are late for work every day; you only reply to your emails in time when it's urgent; and in group tasks, you are always the last person finishing the assigned work!" In other words, your friend notices that your behaviors suggest that you have a strong implicit attitude that seems to have the content like, "my job is bad," even though you may honestly say that your job is good.[5] According to BEL, the information your colleague provides about your behaviors should not influence your previous judgment at all, since it has nothing to do with whether the job *itself* is good. However, it does seem in this case, that you

[3] To follow BEL a person need not be aware of the rule, or even to have a representation of the rule, "…one may follow a rule without realizing that this is what one is doing. Indeed, presumably many non-human animals are permanently in this predicament" (102).

[4] Byrne does notice that the resulting belief of BEL does not always amount to knowledge, however, he does think that BEL is strongly self-verifying, that the resulting belief must be true (112, 107).

[5] Normally we say a person is honest when his/her assertion matches his/her beliefs. This is not what I mean here: I take a negative characterization for honesty in this paper, that one is honest when he/she is not trying to deceive or conceal information in any way. I will talk about this qualification in section 4.

don't actually believe that your job is good, and you are just not aware of that. Two problems emerge from this example: first, it seems there is a possibility that when a person S tries to follow BEL, S does not need to believe that the antecedent is true. Second, it seems there is a possibility that BEL is not strongly self-verifying because the resulting belief is not guaranteed.

Cases of implicit biases are not too far away from this case of your implicit attitude toward your job. First, we can have implicit biases not only toward social groups, but also a wide range of ordinary concepts.[6] Second, there is a significant amount of empirical data showing that it is common that people's explicit attitudes and implicit biases diverge. Nosek et al. (2007) reviewed "2.5 million completed IATs and self-reports across 17 topics obtained between July 2000 and May 2006," and part of their conclusion is that for 5 topics, people's implicit biases correspond only moderately/weakly with their explicit attitudes (2, 42). And if for many of these cases, similar to the job case, our implicit attitudes actually track our beliefs, then BEL might be in trouble. This is my route of argument.

In this paper, I utilize both philosophical writings and empirical findings in psychology, to argue against the normative part of Byrne's account. In other words, although I agree that BEL may work well for ordinary beliefs like "it's raining now" or "there's an apple on the table", it fails to be a good *general* rule because it fails in many cases of implicit biases. I first set the stage of specifying what counts as "trying to follow the rule" (section 2). I then show that many cases of implicit biases are beliefs by arguing that they have many characteristics that are paradigmatic of beliefs (section 3). Then I argue that BEL not only fails as a way to discover many of our implicit biases (section 4)—it does not inform us about what we believe; it also gives the wrong verdict for many cases of implicit biases—it informs us that we believe what we

[6] For example, we can have implicit biases toward vegetables and meat (De Houwer & De Bruycker, 2007), physical activity (Muschalik, 2018), or a specific person (Van Dessel et al. 2018).

do *not* believe (section 5). So BEL cannot be a good general rule for obtaining self-knowledge of our own beliefs. Next, I deal with two possible responses Byrne could offer in defense of BEL (section 6). Finally, I gave a possible explanation for why BEL is not good by utilizing the theory of the left interpreter and discuss the philosophical implications of the theory of the left interpreter (section 7).

2 TRYING TO FOLLOW BEL AND HONESST ASSERTIONS

I'm going to set the stage in this section by arguing for a definition of "trying to follow the rule" in terms of honest assertions.

As part of the descriptive claim, Byrne's notion of "trying to follow the rule" is insufficient to capture how people use BEL. As I mentioned in the introduction, Byrne allows two ways of utilizing BEL concerning a proposition p: a person S follows BEL just in case she knows that the antecedent obtains, therefore believes that she believes that p; a person tries to follow BEL just in case she believes that the antecedent obtains, therefore believes that she believes that p. However, as the job example shows, there might be a third possibility when S tries to follow BEL, in which S does not believe that the antecedent obtains, and yet she comes to believe that p by trying to follow BEL. To include this possibility, I suggest the following definition for "trying to follow the rule" in terms of honest assertions:

> A person S tries to follow BEL iff she could honestly assert that *p*, and then she believes that she believes that *p*.

S tries to follow BEL just in case if she were to ask herself, "is it p?" she could honestly answer that "yes, p." Then she comes to believe that she believes that p. I think this way of capturing "trying to follow BEL" is better than Byrne's because, from a first-person perspective,

S can only be certain that she has an honest assertion. If she were to know or believe that she believes that p or that she knows that p, then there is no need for applying the rule since, according to Byrne, the rule only guarantees true belief, which S already possesses in this situation.

3 IMPLICIT BIASES ARE BELIEFS

After defining "trying to follow the rule" by honest assertions, in this section, I show that there are strong reasons for us to think that some implicit biases are beliefs. And I will do so by arguing that implicit biases can satisfy four characteristics that are paradigmatic of beliefs. We will see in the next section that, in many cases of implicit biases are beliefs, then there is a huge problem for the claim that BEL is strongly self-verifying.

3.1 WHAT ARE BELIEFS

First some background about implicit biases. Implicit biases like implicit racial biases or implicit sexism are pervasive.[7] However, people rarely think that they are racists or sexists. Here's a detailed daily-life example in Schwitzgebel (2012) to display what kind of beast implicit biases is:

> Many men in academia sincerely profess that men and women are equally intelligent. Ralph, a philosophy professor let us suppose, is one such man. He is prepared to argue coherently, authentically, and vehemently for equality of intelligence and has argued the

[7] See Nosek et al. 2007, which reports the result of the online IAT test (https://implicit.harvard.edu/implicit/) that had more than 2.5 million subjects completed the test by the time of this report. This website only contains IAT for implicit biases toward different social groups (people in different races, gender, age, etc.).

> point repeatedly in the past. And yet Ralph is systematically sexist in his spontaneous reactions, judgments, and unguarded behaviors. When he gazes out on the class the first day of each term, he cannot help but think that some students look brighter than others – and to him, the women rarely look bright. When a woman makes an insightful comment or submits an excellent essay, he feels more surprised than he would were a man to do so, even though his female students make insightful comments and submit excellent essays at the same rate as his male students. When Ralph is on the hiring committee for a new office manager, it will not seem to him that the women are the most intellectually capable, even if they are; or if he does become convinced of the intelligence of a female applicant, it will have taken more evidence than if the applicant had been male. And so on.

Cases like Ralph are the focus of my paper, where people exhibit behaviors that reflect certain biases while they are unconscious about it, although specific behaviors vary from subjects to subjects, and the person need not be explicitly rejecting the bias. The experimental results for whether implicit biases in general track people's behavior are mixed[8]; however, in specific topics, implicit biases are shown to be predictive of people's behaviors. For example, Saul (forthcoming) argues that cases of implicit sexism like Ralph's are not rare in the field of philosophy. Hall et al. 2015 reveals that the implicit racial bias of healthcare providers is significantly related to patient-provider interactions, treatment decisions, treatment adherence, and patient health outcomes. As for the methodology, most scientists use the IAT, implicit association test, as the test for implicit biases.[9] Although the test is far from a perfect indication

[8] Greenwald et al. 2009 showed that implicit biases are predictive to behaviors, while a recent review Forscher et al. 2019 gives the opposite result.

[9] See Greenwald 1998. During the test, the subjects will be asked to do sorting tasks in a very short time. For example, a paradigmatic test for implicit bias of sexuality and family vs. career is to pair gender-related names with family or career-related words. The idea is that people's speed and accuracy of sorting the words are indicative to their implicit attitudes, in this case their

of people's implicit attitudes, it's the most commonly used test in rigorous scientific studies, so I will assume that IAT is a relatively reliable test for people's implicit attitudes.

3.2 WHY MANY IMPLICIT BIASES ARE BELIEFS

Mandelbaum (2016) argues that implicit biases are beliefs by showing that they are reason-responsive, that they can be revised in response to counter-evidence. I concur with Mandelbaum's theory but I give additional arguments to prove this claim. More specifically, I argue that including reason-responsiveness, implicit biases exhibit four characteristics that are paradigmatic for beliefs, which are designed to distinguish belief from other attitudes, i.e., imagining.[10] I wish to convey the feeling that these characteristics are paradigmatic for beliefs and focus on why implicit biases may satisfy all of them. Here are the four characteristics:

1) Beliefs respond to evidence. Your belief that P would most likely change if you encounter evidence supporting ~P.

2) Beliefs guide behaviors across contexts. More specifically, across contexts, beliefs guide both unconscious automatic behaviors and conscious behaviors.

3) Beliefs govern other cognitive attitudes. That is, beliefs provide the basis for the formation of other cognitive attitudes like imagining. More specifically, beliefs are the informational background for the formation.

4) Beliefs are involuntary. You cannot just instantaneously choose to believe something or choose to stop believing something.

---

attitude toward the relation between male/female with family/career. If one has low accuracy pairing female-related names with work-related words and pairing male-related names with family-related words, but has high accuracy pairing female-related names with family-related words and pairing male-related names with career-related words, the person may have an implicit bias to associate female with family and male with career.

[10] These characteristics are adapted from Van Leeuwen *forthcoming*. His terminology for beliefs is "factual belief".

3.2.1 Reason-responsiveness

Beliefs respond to evidence, in the sense that when facing reliable counter-evidence, your beliefs will most likely be revised. I would stop believing that it's raining when I see that the rain stops. I would stop believing that there are beers in the refrigerator if I open the refrigerator and find that it's empty. On the other hand, knowing that what you imagine is not true would not stop you from imagining it. Acceptance is another type of attitude, and you can accept an assumption even when you know that the assumption is false, i.e., accepting a simplified situation to focus your attention on a particular issue.

Evidence for the reason-responsiveness of implicit biases is mixed. Mandelbaum (2016) argues that implicit biases can be reason-responsive, which means that they can be revised in the face of counter-evidence. For example, Gregg et al. (2006) first induced subjects' implicit biases toward two fictional tribes, one positive and one negative. Then the subjects were given mere verbal instruction to exchange the characteristics of the two tribes, and IAT shows that their implicit biases are weakened (Mandelbaum, 17). However, Levy (2015) points out that there are also experiments showing that implicit biases are not as reason-responsive as normal beliefs. For example, in Han et al. (2006), subjects were given detailed information about two Pokémon characters. They were then shown a video interviewing small children expressing their attitudes toward Pokémon characters that counters what they've learned. Study shows that even though the subjects rate the children's opinions as not reliable, their IAT shows consistency with the attitudes adopted by the children, rather than the information they previously learned. If implicit biases are strongly reason-responsive, since the subjects rate the new evidence as unreliable, we

should expect their IAT result not reflecting the new evidence. The upshot is that implicit biases can be reason-responsive but are not as reason-responsive as paradigm beliefs.

3.2.2 Guiding Behaviors Across Context

Beliefs guide both conscious and unconscious behaviors across contexts. Conscious behaviors usually are results of one's deliberation on how to attain a certain state of the world, and beliefs guide every step of the deliberation because it provides us an informational background of the world. For example, I want to quench my thirst. I remember that there is beer in the fridge so I walk to the fridge, take the beer and drink it. This deliberation depends extensively on my beliefs: I need to know what a fridge is, what beer is, and that drinking beer can quench my thirst, etc. And beliefs guide your behavior like this no matter what context you are in. On the other hand, beliefs also guide unconscious, fast, automatic actions like the ones tested by the IAT. Greenwald et al. (2009) show that when the IAT and the explicit measure are in alignment, both tests' predictability arises.[11] When explicit attitudes predict behaviors, they are very likely beliefs. So the results suggest that beliefs guide unconscious behaviors. On the other hand, non-belief attitudes like imagining and acceptance guide behaviors only in specific contexts.[12] A wooden stick is a steel sword for a child in one war game and becomes a magical wand in another game. As for acceptance, for example, you would accept that you cannot kick the ball only in basketball, not soccer.

Implicit biases guide unconscious, automatic behaviors, since this is how we come to identify implicit biases. As I mentioned in 3.1, implicit biases also guide people's conscious behaviors.

[11] "Predictability" is about predicting "a wide variety of… physical action, judgments, preferences expressed as choices, and physiological reactions" that is not tested by the IAT test (19).
[12] See Bratman 1992 for context dependence of acceptance. I will also mention this in section 6.2.

Except for studies mentioned in 3.1, Agerstrom and Dan-Olof (2012) find that people's implicit bias against obesity predicts interview decisions. More specifically, they "found a strong and consistent relationship between hiring managers' automatic antiobesity bias and the probability that they would invite an obese, but not a normal-weight, job applicant for an interview…Moreover, the automatic obesity stereotypes predicted discriminatory hiring decisions above and beyond explicitly endorsed hiring preferences and obesity stereotypes. In fact, the latter did not significantly predict hiring decisions" (797). A plausible explanation is that, similar to beliefs, implicit biases guide our deliberation and actions. The manager has the goal of hiring a person suitable for the job. When he is viewing an application with a photo of an obese person, his implicit bias may lead him to doubt the various abilities exhibit in the application, which eventually results in rejecting more obese applicants.[13]

3.2.3 Governing Other Cognitive Attitudes

Beliefs govern other cognitive attitudes, that is, beliefs are the basis for the formation of other cognitive attitudes. Van Leeuwen (2013) provides a nice illustration for beliefs' governing imaginings:

> …imagine a female lion in the bushes looking out on a herd of grazing antelope; then imagine she charges them. What do you imagine next?
>
> I'm guessing you imagine the antelope run away. But why? Imagining this wasn't in the instructions. You could have imagined them staying put, though you didn't. The answer, around which some consensus has emerged, is that one's beliefs inferentially

[13] Similarly, Dovidio& Gaertner 2000, Uhlmann& Cohen 2005 and Son Hing et al. 2008 also show that implicit bias influence people's hiring behaviors.

> govern transitions from initial imagining to later imaginings. The proposition *a lion charges antelope* does not by itself entail the antelope run away. But if we add *antelope flee charging lions*, which you believe, then we get the entailment. (page 8, Van Leeuwen's italicization)

Like beliefs, implicit biases also govern other cognitive attitudes. For imagining, if you were to give IAT tests to Ralph, he would perform better when grouping male faces with good words like intelligent, diligent, smart, etc. So it's reasonable to assume that when Ralph is asked to imagine a good student, the first figure that comes to his mind would very likely be a male student. Similarly, try to picture in your head the following people and give yourself several seconds for each one: a drug dealer, a genius, a good parent, and a successful businessperson. And notice how the people you imagine may have different races and sexes. These imaginings may be results of your implicit biases, or some weaker implicit attitudes. Either way, it's reasonable to presume that implicit biases have influences like this to your imagining. Implicit biases also govern inferences. For example, Adam has a strong implicit bias about male and female and their relationship with career and family. He is much more accurate when pairing male-related names with career-related family. When he tries to reason about what is important for his life, even though he might be explicitly thinking that it's ok for men to focus on family, he may simply not feel right to only focus on family for *his* life, and he may justify this feeling by providing various rationalizations. If Adam's implicit bias is strong, we would observe that that inference like this would be common in his life, and this will be the case especially when he is not being self-monitoring or he is self-monitoring but does not want to change.[14]

[14] Payne 2005. I will return to this point in section 6.2.

3.2.4 Involuntariness

Most philosophers take beliefs to be involuntary.[15] I cannot simply form a belief that P instantly just because I want to. For example, I cannot just instantly stop believing that today is Wednesday because my paper is due on Wednesday and I haven't finished it.[16] On the other hand, attitudes that are not beliefs, like imagining or acceptance, are under direct voluntary control to some extent. You can instantly choose to imagine a pear rather than an apple, or to accept different propositions as the basis of your deliberation.

Implicit biases are similarly involuntary. You cannot just decide not to have, or to develop a new, implicit bias and it will instantly disappear or appear. If one day Ralph realizes that he has the implicit bias against people of other races, he cannot abandon that implicit bias by simply deciding not to have it. His automatic behaviors will continue to show that he has the implicit bias.[17]

So far, we've seen that implicit biases exhibit almost all major characteristics of beliefs: implicit biases can be reason-responsive to some extent; they govern our behaviors across contexts; they govern the formation of other cognitive attitudes; and that they are involuntary. Even though implicit biases may not be perfectly reason-responsive, first, it is not the case that all beliefs in all situations are perfectly reason-responsive. Suppose that Adam falls in love with Juliet, and Juliet is nice to him. So Adam believes that Juliet likes him. Even when there is counter-evidence, that Juliet also shows affection to another man, Adam's belief would likely

---

[15] Even for philosophers who support direct voluntarism, they acknowledge that the voluntariness is very limited (e.g., see Ginet 2001). Here I assume that direct doxastic voluntarism is false.

[16] Notice that we still have indirect voluntary control to our beliefs, that our beliefs can be changed after we take some intermediate actions. For example, if I have a belief that it's raining now and I don't want to have it because I'm dying for camping, I could search on the Internet and gather more information about today's weather and probably find credible source predicting that the rain will stop in an hour, in which case my belief may revise.

[17] As normal beliefs, this doesn't mean you cannot change your implicit biases through *indirect* voluntary control. For example, you could give yourself processes of conditioning to weaken your implicit biases (Kurdi& Banaji, 2019). Or, as we saw in Mandelbaum (2016), you could also weaken your implicit biases by encountering counter evidences on the subject of your implicit biases.

remain (even though the strength of the belief might be weakened). Second, implicit biases satisfy all other characteristics of beliefs well, especially that they are the basis for the formation of other cognitive attitudes and that they govern our behaviors across contexts. Therefore we have a strong case for the claim that some implicit biases are beliefs. And so, if using BEL cannot discover some of these implicit biases or even gives the wrong verdict for some of them, then BEL is not a good rule. I will talk about these two kinds of cases in turn.

4 FAILURES OF DISCOVERING IMPLICIT BIASES

People with implicit biases are not all like Ralph, in the sense that they do not all explicitly endorse propositions contrary to their implicit biases. Suppose Adam is another person with roughly the same implicit bias against women that Ralph has. Now someone goes to Adam and asks him whether he thinks he has the belief that women are inferior in intelligence to men (let's call this proposition, *P*); there are three possibilities. First, Adam explicitly believes that that *P*. Second, he is explicitly ambivalent about *P*. Third, like Ralph does, he explicitly rejects *P*. The first situation is not a problem for Byrne because remember that BEL is:

BEL: If *p*, believe that you believe that *p*.

Also, remember that to follow BEL means to conclude that one has the belief that *P* when one thinks that *P.* So if Adam explicitly believes that *P,* he could successfully move from *P* to the conclusion that he believes that *P*. However, the other two kinds of cases are problematic. I call the second case where Adam is explicitly ambivalent about *P*, cases of missing answer, which will be the focus of this section. I call the third case where Adam explicitly rejects *P*, cases of wrong verdict, which will be the focus of my next section.

In the second case, Adam is explicitly ambivalent about *P*: "I think I need more empirical evidence to believe either way." When he tries to apply BEL, because he is ambivalent about *P*, he cannot think that *P* so he cannot move to the conclusion that he has the belief that *P*. Therefore, Byrne's rule fails in this case because it's missing an answer.

However, of course, there are ways that one could know about one's belief that *P*, namely by self-monitoring, observing one's own behaviors, and maybe listening to other's evaluations toward oneself. This is exactly what you should do if you are in the situation described in the introduction: you should take what your colleague says into your considerations for whether you have the belief. Notice, however, these methods of acquiring self-knowledge of belief through investigating one's own behaviors are exactly not the method provided by BEL. To follow BEL, one can only ask oneself the question "Is P true?" then deliberate on it and offer an answer. To make things clearer, following Moran (2001)'s terminology, there are two approaches for tackling the question "Do you think that P?": taking a theoretical stance toward this question is to view my belief as an object of discovery, as if I were to discover whether a piece of metal is iron. That is, you would observe your behavior and the mental states to investigate whether you have the belief, like you would perform chemical tests on this piece of metal and see what reaction would happen (Moran, 63). On the other hand, we can also take a deliberative stance toward this question, which is to understand that this question is about *your* belief, and you would have this belief just in case *you* think that P is true. This is exactly the transparency character that Byrne wishes to capture by BEL. You simply need to consider, "Is it P?" If your answer is yes, then you have the belief that P. So it should not be surprising that taking a theoretical stance toward the question is not following BEL. Therefore, the cases of missing answer remains to be a problem to Byrne's evaluative claim.

5 CASES OF WRONG VERDICT

The third kind of cases that I call cases of wrong verdict creates a bigger problem for BEL. Compared to the second kind of cases, where Adam is ambivalent about *P*, in the third kind of cases, he explicitly thinks and asserts that the opposite of *P* is correct: "Women certainly have the same level of intelligence as men!" Similar to the missing answer cases, since he does not explicitly think that *P*, he cannot apply BEL and acquire the belief that he believes that *P*. Furthermore, it seems that Adam can even get the opposite answer for his belief if he applies BEL to check whether he has *not P*. "Is it the case that men and women are equal in intelligence? Yes. According to BEL, I must, therefore, believe that men and women are equal in intelligence." Hence he would acquire the belief that he believes that *not P*, which is the opposite of his actual belief. I call these cases, cases of wrong verdict.[18]

As I mentioned in the introduction, empirical data shows that it is common that people's explicit attitudes and implicit biases diverge. So if my account is correct that many of these implicit biases are beliefs, then all of these cases would be cases of wrong verdict where BEL fail, thus creating a huge problem for Byrne's normative claim that BEL is good.

6 REPLIES TO BYRNE'S TWO POSSIBLE RESPONSES

Here is a quick recapitulation of my argument so far. Many implicit biases are beliefs because they largely satisfy four characteristics that are paradigmatic to beliefs: that they are involuntary, they guide people's non-verbal actions, they govern other cognitive attitudes, and

[18] Notice that it seems possible that there is another kind of cases where BEL gives a wrong verdict, those are cases where although the subject explicitly endorse *P*, he has no beliefs in either *P* or not *P*. So when the subject applies BEL to *P*, he would acquire the false belief that he believes that *P*. Since I'm focusing on cases of implicit biases, these cases can be a future topic.

they are partially reason-responsive. Byrne's theory argues that the epistemic rule for belief, BEL, that if *P*, believe that you believe that *P,* is a good rule. If my account for implicit biases is correct, however, there could be at least two kinds of cases where BEL seems to be inapplicable: cases of missing answer (the subject is explicitly ambivalent about the content of the belief and so BEL would miss this belief) and cases of wrong verdict (the subject explicitly asserts the opposite of the belief and so BEL would give an answer that is the opposite of the actual belief).

In this section, I will introduce two possible responses from Byrne to my objections and offer my replies to them.[19] The first possible response is that Byrne could argue that some characteristics, like transparency or peculiar access, are necessary for beliefs.[20] Implicit biases are not beliefs, since they fail to have these characteristics. Byrne's second possible response is that, even if we grant that implicit biases are beliefs, it is unclear that BEL is giving the wrong verdict in the second case. It seems reasonable that people can hold contradicting beliefs, and what BEL picks out is only the conscious belief that *not P*. In other words, BEL is only missing some of the beliefs, but is not providing wrong information about them. I don't think either of Byrne's possible responses would work.

6.1 Reply to Byrne's First Possible Response

For the first response, by saying that some characteristics are necessary for beliefs, one is either making a conceptual claim or an empirical claim. If it is a conceptual claim, then it's saying that beliefs are defined partly by such characteristics, and the claim would not be very

[19] Because there are several layers to this discussion, I call my main argument which objects to Bynre's BEL rule, "objection"; I call these two possible comments from Byrne to my objection, "responses"; Finally, I call my replies to these responses, "replies".

[20] For Byrne, briefly, having peculiar access to our mental states means that we can access our mental states in ways that others cannot (8).

interesting, because we do think that what beliefs are is not only a matter of stipulated definition. There is a fact of the matter about the nature of belief, which we are trying to discover. On the other hand, ever since Freud (or maybe even earlier), philosophers acknowledge the existence of unconscious beliefs that significantly influence people's actions. Moran, for example, in his 2001 book, discusses cases of beliefs that may not be transparent (Moran, 85). A belief *P* is transparent just in case I can conclude that I believe *P* when I think *P* is true. If it were a conceptual truth that beliefs are transparent, then these philosophers would be just making conceptual mistakes.

The second interpretation is that when one says beliefs necessarily involve …, one is making an empirical claim. If so, one is saying that our current empirical evidence strongly suggests that all beliefs have such characteristics. However, Byrne did not provide enough empirical evidence to show this. For peculiar access, he only quotes the psychologist Wilson to support his claim: "I can bring to mind a great deal of information that is inaccessible to anyone but me. Unless you can read my mind, there is no way you could know that a specific memory just came to mind" (Wilson 2002:105, Byrne 11, 12). However, if implicit biases govern other cognitive attitudes like imagining, we also have some peculiar access to implicit biases, namely we could directly monitor the formation of our cognitive attitudes and check if it consistently implies the existence of an implicit bias, and no one else is able to do this. On the other hand, it's true that implicit biases are largely not transparent, but my point is exactly that, because they do largely exhibit many characteristics that are paradigmatic of belief, we should count them as beliefs. Therefore, Byrne's first response is not satisfactory.

6.2 Reply to Byrne's Second Possible Response

Byrne's second possible response is that, people could hold contradicting beliefs. People could have an unconscious belief that *P*, while consciously believe that not *P*. Going back to our example, Ralph could have an unconscious belief that women are inferior to men in intelligence, while having the explicit belief that men and women have the same level of intelligence. When he applies BEL, he thinks that "women are *not* inferior to men in intelligence; therefore, I believe that women are *not* inferior to men in intelligence," and this application of BEL would be successful, because he really has this explicit belief.

My initial reply is that, for the two kinds of problematic cases that I mentioned (the cases of missing answers and the cases of wrong verdict), if this second response of Byrne works, it only works for the cases of wrong verdict. Because even though BEL won't give the wrong verdict, it still cannot discover the unconscious belief that *P*.

On the other hand, this second response doesn't work for the cases of wrong verdict, either. I admit that there could be cases where although the implicit bias governs many of the non-verbal behaviors and the formation of cognitive attitudes, the explicit attitude also plays similar functional roles, in which case we could arguably call both attitudes belief (or neither). My focus, however, is the significant amount of cases of NO AWARENESS:

> **NO AWARENESS**: a case is a NO AWARENESS when the person is not aware of his/her having the implicit bias that *P,* and his implicit bias governs almost all of their relevant non-verbal behaviors and cognitive attitudes formation, although they honestly endorse *not P*.

I've been avoiding the problem of what "honestly" means in these situations, that is, what attitude people have when they only explicitly endorse *not P*. I now argue that in cases of NO

AWARENESS, there is only one belief, the implicit bias. The format of my argument is similar to that in section 2: explicit attitudes in these cases do not exhibit many of the characteristics that are paradigmatic to beliefs, so we shouldn't count them as beliefs.

In cases of NO AWARENESS, explicit attitudes do not govern non-verbal behaviors. Even though Ralph explicitly endorses that *not P*, without awareness of his implicit bias, his various non-verbal actions would consistently exhibit his implicit bias. Payne (2005) examines the relationship between levels of self-control and the divergence of implicit and explicit attitudes and found that "for people with a score of 0 in control, stereotypical errors are a one-to-one reflection of automatic bias" (496). Also, "Participants with a strong automatic bias formed a more negative impression of the target character, but this was especially so if they were also low in cognitive control" (499). Payne's study supports my claim because when you are not aware that you have the implicit bias that *P*, you cannot have self-control over it, in which case Payne's result suggests that you would exhibit stereotypical errors and form negative impressions of the target character. This implies that your implicit bias would be guiding these non-verbal behaviors rather than your explicit attitudes.

In cases of NO AWARENESS, explicit attitudes also do not (largely) govern other cognitive attitudes. Since Ralph is not aware of his having the implicit bias and thus does not intend to control it, it's plausible to assume that the implicit bias would govern the formation of his cognitive attitudes, like the example of imagining a good student. The point is more difficult to make when it comes to explicit reasoning. There are clear cases where explicit attitudes do govern explicit reasoning, for example, the verbal endorsement itself. However, it seems that this governance is there only when the person is consciously aware that the subject of inference is related to the content of the explicit attitude. When Ralph is consciously deciding which

candidate is better for the job, although this deliberation is objectively related to his explicit attitude that women are as intelligent as men, if it doesn't come to her that the subject matter relates to this explicit attitude, the explicit attitude will not play a role in this deliberation. In contrast, implicit biases always play a role in inferences. Even in conscious reasoning where Ralph makes use of her explicit attitude, in a very simplified way, he has to overcome his implicit biases.

7 A POSSIBLE EXPLANATION FOR THE FAILURE OF BYRNE'S ARGUMENT

A quick summary of what we have done so far: Byrne offers a descriptive theory of self-knowledge that we gain self-knowledge by following an epistemic rule, BEL.

BEL: If *p*, believe that you believe that *p* (Byrne, 103).

Byrne further argues for the normative claim that BEL is a *good* rule, in the sense that if one tries to follow the rule, the resulting belief is guaranteed to be true. My strategy so far is to argue for the existence of a nonnegligible amount of counterexamples for Byrne's normative claim, that for many cases of implicit biases, BEL would either miss or give a wrong answer concerning our beliefs. So BEL is not a good rule.

I have not directly argued against Byrne's argument for the goodness of BEL, because Byrne basically defines "trying to follow BEL" in the way that analytically, the resulting belief is true. As I mentioned in sections 1 and 2, Byrne defines "trying to follow BEL" as first believing that the antecedent obtains, then believes that one believes that *p*. If this is the definition of "trying to follow the rule," then a person trying to follow BEL analytically implies that the person believes that the antecedent obtains, which means that the resulting belief is true. However, as I argued in

section 2, a better way of capturing "trying to follow the rule" is by defining it in terms of honest assertions. If this is the case, then Byrne's underlying argument is this:

P1. One trying to follow BEL with regard to *p* implies that she could honestly assert that *p*, more specifically, if she asks herself whether *p* is true, she would honestly reply yes.

P2. Being able to honestly assert that *p* implies having the belief that *p*.

Therefore,

C. One trying to follow BEL with regard to *p* implies that she believes that *p*.

If this were his underlying argument, then my strategy is also a refutation to his argument because it shows that P2 is false. Cases of implicit biases that I mentioned are counterexamples to P2 because these are cases where people could honestly assert that *p* but do not believe *p*. Ralph could honestly assert that women are equal to men in intelligence, even though he does not actually believe so.

So what is left for me to make my argument more intelligible is to give a possible explanation for why P2 is false. That is, why are the counterexamples, where people honestly assert that *p* but do not believe that *p*, possible. I will appeal to the possible underlying cognitive mechanism that enables counterexamples of P2, the mechanism of the left-brain interpreter.

Michael Gazzaniga, one of the founders of the field of cognitive neuroscience, has been conducting split-brain research for more than 50 years now, producing numerous books and articles. Split-brain patients are patients whose corpus callosum that connects the two halves of the brain is severed, to the extent that there is barely any information communication between the two halves. One of the theories that he offers through experiments with split-brain patients is the theory of the left-brain interpreter, that our left brain controls our assertions and provides us with the best stories given the information that it receives (Gazzaniga 2000, 1316; Gazzaniga

2011, 83; Gazzaniga 2018, 211). According to Gazzaniga, here are some relevant facts about a typical split-brain patient: 1). “any visual, tactile, proprioceptive, auditory, or olfactory information that was presented to one hemisphere was processed in that half of the brain alone, without any awareness on the part of the other half” (Gazzaniga 2011, 57). For example, if a picture is only presented to the patient’s left visual field and not the right visual field, then only the half of the brain that processes information in the left visual field can acquire the content of the picture, which means that only the right hemisphere can “see” the picture. 2). The left hemisphere controls most of our linguistic abilities, while the right hemisphere has a very limited speaking ability. When a picture is presented only to the left visual field, the patient can barely describe anything. While if the picture is presented only to the right visual field, then the patient can appropriately describe it (56). 3). Both hemispheres are able to read simple words and relate words to pictures. 4). The left hemisphere is good at making inferences, while the right hemisphere is bad at it (62).

Gazzaniga observed the following phenomenon:

> We showed a split-brain patient two pictures: A chicken claw was shown to his right visual field, so the left hemisphere only saw the claw picture, and a snow scene was shown to the left visual field, so the right hemisphere only saw that. He was then asked to choose a picture from an array of pictures placed in full view in front of him, which both hemispheres could see. The left hand pointed to a shovel (which was the most appropriate answer for the snow scene) and the right hand pointed to a chicken (the most appropriate answer for the chicken claw). Then we asked why he chose those items. His left-hemisphere speech center replied, “Oh, that’s simple. The chicken claw goes with the chicken,” easily explaining what it knew. It had seen the chicken claw.

> Then, looking down at his left hand pointing to the shovel, without missing a beat, he said, "And you need a shovel to clean out the chicken shed" (82).

Gazzaniga notices two interesting points from this experiment. First, even though the information about the picture in the left visual field cannot be delivered to the left hemisphere, rather than replying that it cannot explain the movement of the left hand, the left hemisphere gave a relatively coherent story with the information at hand. Second, notice that the patient is not consciously making the rationalization; he does think that he points to the shovel because of the reason he just provided. In other words, the patient is making *honest assertions* when he is making the claim. In support of this, Gazzaniga describes that "…without batting an eye, it (the left interpreter) would incorporate the right hemisphere's response into the framework…the left interpreter did not offer its suggestion in a guessing way but rather as a statement of fact as to why that card has been picked" (Gazzaniga 1978, 148). Gazzaniga finds that various other experiments with split-brain patients consistently show this effect, that their left brain would provide the best explanation for the behaviors controlled by the right brain with the information that it receives, and so he names this faculty of interpreting, the left-brain interpreter, or the left interpreter (Gazzaniga 2000, 1318).

Although split-brain patients are different from normal people, Gazzaniga's experiments did show that our left brains have this ability of an interpreter, the ability to make up stories as coherent as possible, and that the accuracy of the story depends greatly on how much information it could receive. Moreover, since the whole process is unconscious, the person can honestly assert the story that her left hemisphere came up with.

This theory of left-brain interpreter provides a possible explanation for why P2 is wrong, that our honest assertions do not always track our underlying beliefs: first, the story that the left

interpreter presents largely depends on the information that it receives. Normally the interpreter could receive the relevant information from other parts of our brain, which makes our assertions accurate most of the time. However, it could be that, due to some reasons (probably psychological), the left-brain interpreter cannot access all the relevant information regarding some of our beliefs. In this case, the story it provides would most likely not correspond to the beliefs.

Second, it is not clear that the left interpreter solely aims at truth, in the sense that the evolutionary function of the interpreter is to track the truth. On the one hand, the left interpreter trades accuracy for coherence. For example, in experiments with the split-brain patients, if you show the pictures of getting up in the morning and making cookies to one of the hemispheres and ask the person to identify those pictures among a new series of pictures whose content is not related, both hemispheres could identify well what pictures they saw. However, if some of the pictures in the new series are related to getting up in the morning and making cookies, then only the right hemisphere could accurately identify the pictures that it saw, while the left hemisphere would incorrectly identify some of the new and related pictures as seen previously (Gazzaniga 1998, 26). The explanation seems to be that because the new pictures fit the pattern, which makes the story more coherent, the left interpreter gladly abandons accuracy for coherence. On the other hand, which might be related to the first point, it seems that the left interpreter aims to preserve a certain complete self-image. This is supported by the fact that the split-brain patients would not notice anything wrong after the surgery that served their corpus callosum. This is quite surprising if you think about it: the left brain has lost all information input about the mental processes going on in the right brain, and yet "…this system still lets us feel like 'us'" (Gazzaniga 2011, 103). If this is correct, then the left interpreter even trades coherency for a

complete self-image, since the true story that the left brain has no idea about what the right brain is doing is more coherent than the made-up story by the left interpreter.

That the left interpreter could trade accuracy for coherence and a complete self-image could explain why the implicit biases rather than beliefs like "it's raining today" are more difficult to obtain through BEL: the implicit biases might be inconsistent with certain self-image. Normally, people have a good self-image in the sense that we are good people. If the person regards the implicit biases as true, for example, that women are less intelligent than men, then the self-image is potentially threatened, since a good person would presumably do not have (irrational) biases against other people. Therefore, to preserve the self-image, the left interpreter presents a story that is coherent enough, but more importantly is consistent with the self-image. This may also explain why the psychiatrists could point out facts about your beliefs and other attitudes that seem to you shocking at first glance.

One may comment that I cannot explain implicit biases with the theory of left interpreter yet, because it seems that we have so far only seen the effect of the left interpreter vividly in the split-brain patients. It seems that normal people do not present other split-brainlike phenomena.[21] My response is that, it is true that in the normal population, we do not observe phenomenon as dramatic as those observed in the split-brain patients. And this is partly predicted by the theory: the story that the left interpreter provides is highly constrained by the information it receives. So it makes sense that the split-brain patient's left interpreter makes a more inaccurate story than a normal person, because it is more constrained in the information it could receive. However, some salient characteristics of the normal population can still be well explained by the left interpreter. Remember that according to the split-brain patient studies, the left interpreter makes up stories

[21] Thanks to Dan Weiskopf for this comment.

according to the information it received that aims for coherence and a consistent self-image. As mentioned above, for split-brain patients, this function is shown, first, in the experiment where the left brain tends to count new images as seen if those images could constitute a more coherent story; second, the patient is entirely ignorant of the situation after their corpus callosum is severed. In the normal population, on the other hand, the theory of left interpreter can explain the long-known superiority illusion, that people view themselves as better in their abilities or character traits than the average person. The illusion is shown consistently in decades of experiments with regard to various aspects of life (Zell et al. 2019). Just to name a few, people regard them as driving better than the average person; they think they are more popular than the average person; they also think that they are more environmentally friendly than the average person (Sevenson 1981; Zuckerman & Jost 2001; Bergquist 2020). The theory of the left interpreter could explain why the false beliefs that we are better than average are hard to update: this story is better than the fact that we are, in most cases, not better than the average person for the left interpreter because it helps preserve a good self-image.

8 PHILOSOPHICAL IMPLICATIONS OF THE LEFT INTERPRETER

If we accept the theory of the left interpreter, then the first lesson for our theories of self-knowledge is a lesson for the normative theory: a good theory of self-knowledge, in the sense that it consistently generates true beliefs, has to rely on not only the honest assertions of the person but also other information that may indicate what beliefs the person holds. Using Moran's terminology that I mentioned back in section 4, only taking a deliberative stance toward oneself is not enough; we also need to take a theoretical stance toward ourselves. In other words, in order to more accurately obtain self-knowledge, we should also treat ourselves as objects of

investigations. For example, it would be more accurate if the theory takes into account the behaviors of the person and the evaluations of others, as I mentioned at the end of section 4. As cases like the job example show, we are not very good at tracking all of our behaviors, especially when the behaviors are performed unconsciously. This means that the information about these behaviors that might shed light on what beliefs we have may not be easily accessible to the left interpreter, and therefore may not be reflected in our sincere assertions. However, if our theory of self-knowledge includes this component, then by trying to follow the theory, we would put more attention to our behaviors, and therefore our resulting belief may be more likely to be true.

It is true that in most cases, a deliberative stance is enough, where we could know about our beliefs by merely asking ourselves whether the belief is true. For example, whether I have beliefs like "it is raining today," or "I have a pencil in my bag." However, these cases also seem to be the cases where we do not need a theory of self-knowledge to know what we believe. We naturally go from "it is raining today" to "I think that it is raining today" without realizing that we have done anything substantial. So in this sense, the deliberative aspect of the theory of self-knowledge works more as a descriptive claim than a normative claim, in the sense that it accurately depicts what we fail to notice when we acquire self-knowledge in these majority of cases, but it does not guide us to obtain self-knowledge more accurately because we are already following it.

However, if we take the theory of left interpreter seriously, there is also a lesson for our descriptive theory of self-knowledge: an accurate theory of self-knowledge, in the sense that it best explains how we acquire self-knowledge, cannot only have the deliberative stance. A theoretical stance toward oneself is necessary, where the person is treated as an object of investigation. The reason is that whenever we take a deliberative stance toward ourselves, we are

relying on our honest assertions, which are simply stories that the interpreter made up, and the way the left interpreter works seem to rely on taking a theoretical stance toward oneself. Since the corpus callosum is severed and there is barely any information communication between the two hemispheres, in all these experiments, the left interpreter of the split-brain patients made up stories mainly based on the behaviors of herself, or more specifically, behaviors that the right hemisphere dictate. In one experiment, five slides with two words on each slide is shown to the patient (Mary + Ann; May + Come; Visit + Into; The + Town; Ship + Today). The left hemisphere only sees words on the right side while the right hemisphere only sees words on the left. Here is the dialogue between the experimenter (E) and the patient (PS) when the experimenter asks what the patient saw:

> PS: Ann come into town today [*the left hemisphere answers*]
>
> EXPERIMENTER (E): Anything else?
>
> PS: On a ship [*here comes the right hemisphere*]
>
> E: Who?
>
> PS: Ma
>
> E: What else?
>
> PS: To visit
>
> E: What else?
>
> PS: To see Mary Ann
>
> E: Now repeat the whole story
>
> PS: Ma ought to come into town today to visit Mary Ann on the boat. (Gazzaniga 2011, 101. Gazzaniga's italicization)

The interpreter first answers the information it has direct access to, which are the words on the right side, "Ann come into town today." Then, because the right hemisphere has weak ability to utter words, the patient slowly spelled out what is seen by the right hemisphere, and the

interpreter can only put this piece of information into the story after the utterance of the right hemisphere. As Gazzaniga describes, “The interpreter received the information from the right hemisphere externally; it did not have access to this part of the story until it was uttered by the right hemisphere, heard by the left hemisphere, and then the interpreter had to deal with the situation” (101). This shows that even for cases where we think we are only taking a deliberative stance toward oneself by asking, “is it *p*?” The answer, because the interpreter gives it, is already a result involving observing the behaviors of ourselves. Since the interpreter operates at an unconscious level, we still regard our answer as only focusing the fact about *p*. If this is the case, then for any descriptive theory of self-knowledge, a picture that takes only the deliberative stance into account is insufficient.

9 CONCLUSIONS

In this paper, I first argue that implicit biases are beliefs because they exhibit characteristics that are paradigmatic to beliefs, especially that they are involuntary, they guide non-verbal behaviors across context, and that they govern the formation of other cognitive attitudes. Then I argue that since the BEL rule, that if *P*, believes that you believe that *P*, fails to discover a lot of implicit biases and gives wrong verdicts about many implicit biases, it is not a good rule. Next, I dealt with two possible responses from Byrne, the upshot being, in NO AWARENESS cases, where people are not aware of their implicit biases and their implicit biases guide their non-verbal behavior and govern the formation of their cognitive attitudes, their explicit attitudes are not beliefs because it fails to guide behaviors and fails to govern cognitive attitudes across contexts.

Finally, I introduce Gazzaniga's theory of the left interpreter as a possible explanation for why honest assertions are not enough for beliefs. Our left interpreter provides our honest assertions, yet the story it tells depends on how much information it receives, and it does not aim to tell an accurate story. Rather, the left interpreter aims for a coherent story that preserves some self-image, which is exactly what implicit biases cannot contribute to and why the implicit biases are not part of the story. The implication is that any descriptive or evaluative theory of self-knowledge should partly consist of a theoretical stance toward oneself.

Although BEL is intuitive, attractive, and we may actually have this epistemic rule built into our cognitive architecture, we should be careful when taking its verdict as true and always try to be reflective of our judgments. When trying to discover what beliefs we have, in order to reach a more accurate view, we should not only consider our honest assertions, but also take into account of what our behaviors imply about us and how other people think you would believe. So going back to the example in the introduction, you shouldn't rush making the important decision. Rather, you should listen carefully to what your colleague says and take it into serious consideration. You might experience a certain level of cognitive dissonance while you are considering it seriously, but you could avoid potential regret for not going after a new job in the future.

REFERENCES

Agerstrom, Jens, and Dan-Olof Rooth. "The Role of Automatic Obesity Stereotypes in Real Hiring Discrimination." Human Resource Management International Digest 20, no. 1 (2012). https://doi.org/10.1108/hrmid.2012.04420aaa.006.

Bergquist, Magnus. (2020). Most People Think They Are More Pro- Environmental than Others: A Demonstration of the Better-than-Average Effect in Perceived Pro-Environmental Behavioral Engagement. Basic and Applied Social Psychology. 10.1080/01973533.2019.1689364.

Bratman, Michael E. "Practical Reasoning and Acceptance in a Context." Mind 101, no. 401 (1992): 1–16. https://doi.org/10.1093/mind/101.401.1.

Byrne, Alex (2018). "Transparency and Self-Knowledge". Oxford University Press.

Dessel, Pieter Van, Yang Ye, and Jan De Houwer. "Changing Deep-Rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi." Social Psychological and Personality Science 10, no. 2 (2018): 266–73. https://doi.org/10.1177/1948550617752064.

Dovidio, John F., and Samuel L. Gaertner. "Aversive Racism and Selection Decisions: 1989 and 1999." Psychological Science 11, no. 4 (2000): 315–19. https://doi.org/10.1111/1467-9280.00262.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Gazzaniga, Michael S., and Joseph E. Ledoux. "The Split Brain and the Integrated Mind."

The Integrated Mind, 1978, 1–7. https://doi.org/10.1007/978-1-4899-2206-9_1.

Gazzaniga, Michael S. The Minds Past. Berkeley, CA: University of California Press, 1998.

Gazzaniga, M. S. "Cerebral Specialization and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?" Brain 123, no. 7 (January 2000): 1293–1326. https://doi.org/10.1093/brain/123.7.1293.

Gazzaniga, Michael S. Whos in Charge?: Free Will and the Science of the Brain. NY, NY: ECCO, an imprint of HarperCollins Publishers, 2011.

Gazzaniga, Michael S. The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind. New York: Farrar, Straus and Giroux, 2018.

Ginet, Carl. "Deciding to Believe." Knowledge, Truth, and Duty, 2001, 63–75. https://doi.org/10.1093/0195128923.003.0005.

Greenwald, Anthony G, Debbie E McGhee, and Jordan L. K. Schwartz. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." Journal of Personality and Social Psychology 74, no. 6 (1998): 1464–80.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity." Journal of Personality and Social Psychology 97, no. 1 (2009): 17–41. https://doi.org/10.1037/a0015575.

Gregg, A., Seibt, B., and Banaji, M. (2006). "Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences". Journal of Personality and Social Psychology 90 (1): 1–20.

Hall, William J., Mimi V. Chapman, Kent M. Lee, Yesenia M. Merino, Tainayah W. Thomas, B. Keith Payne, Eugenia Eng, Steven H. Day, and Tamera Coyne-Beasley. "Implicit

Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review." American Journal of Public Health 105, no. 12 (2015): 2588–88. https://doi.org/10.2105/ajph.2015.302903a.

Han, H. Anna, Michael A. Olson, and Russell H. Fazio. "The Influence of Experimentally Created Extrapersonal Associations on the Implicit Association Test." Journal of Experimental Social Psychology 42, no. 3 (2006): 259–72. https://doi.org/10.1016/j.jesp.2005.04.006.

Hing, Leanne S. Son, Greg A. Chung-Yan, Leah K. Hamilton, and Mark P. Zanna. "A Two-Dimensional Model That Employs Explicit and Implicit Attitudes to Characterize Prejudice." *Journal of Personality and Social Psychology* 94, no. 6 (2008): 971–87. https://doi.org/10.1037/0022-3514.94.6.971.

Houwer, Jan De, and Els De Bruycker. "Implicit Attitudes towards Meat and Vegetables in Vegetarians and Nonvegetarians." International Journal of Psychology 42, no. 3 (2007): 158–65. https://doi.org/10.1080/00207590601067060.

Kurdi, B., & Banaji, M. R. (2019). "Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features". Journal of Personality and Social Psychology, 116(5), 681-703.

Mandelbaum, Eric (2016). "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias". Noûs 50 (3):629-658.

Moran, Richard A. (2001). "Authority and Estrangement: An Essay on Self-Knowledge". Princeton University Press.

Muschalik, Carolin, Iman Elfeddali, Math J. J. M. Candel, and Hein De Vries. "A Longitudinal Study on How Implicit Attitudes and Explicit Cognitions Synergistically

Influence Physical Activity Intention and Behavior." BMC Psychology 6, no. 1 (2018). https://doi.org/10.1186/s40359-018-0229-0.

Levy, Neil. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." Noûs 49, no. 4 (2014): 800–823. https://doi.org/10.1111/nous.12074.

Nosek, Brian A., Frederick L. Smyth, Jeffrey J. Hansen, Thierry Devos, Nicole M. Lindner, Kate A. Ranganath, Colin Tucker Smith, et al. 2007. "Pervasiveness and correlates of implicit attitudes and stereotypes". European Review of Social Psychology 18: 36-88.

Payne, B. Keith. "Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping." Journal of Personality and Social Psychology 89, no. 4 (2005): 488–503. https://doi.org/10.1037/0022-3514.89.4.488.

Svenson, Ola. (1981). Are We All Less Risky and More Skillful than our Fellow Drivers?. Acta Psychologica. 47. 143-148. 10.1016/0001-6918(81)90005-6.

Uhlmann, Eric Luis, and Geoffrey L Cohen. "Constructed Criteria: Redefining Merit to Justify Discrimination." Psychological Science 16, no. 6 (n.d.): 474–80. https://doi.org/10.1111/j.0956-7976.2005.01559.x.

Van Leeuwen, Neil. "The Meanings of 'Imagine' Part I: Constructive Imagination." Philosophy Compass 8, no. 3 (December 2013): 220–30. https://doi.org/10.1111/j.1747-9991.2012.00508.x.

Wilson, T. D. 2002. "Strangers to Ourselves: Discovering the Adaptive Unconscious". Cambridge, Ma: Harvard University Press.

Zell, Ethan & Strickhouser, Jason & Sedikides, Constantine & Alicke, Mark. (2019). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. Psychological Bulletin. 146. 10.1037/bul0000218.

Zuckerman, E., & Jost, J. (2001). What Makes You Think You're so Popular? Self-Evaluation Maintenance and the Subjective Side of the "Friendship Paradox". Social Psychology Quarterly, 64(3), 207-223. Retrieved April 15, 2020, from www.jstor.org/stable/3090112